Store speech, select vocabulary to recognize word

A method for speech recognition through storing a stream of concatenated speech items, and therefrom in successive cycles selecting a non-recognized speech item for recognizing viz à viz an associated limited and selectable vocabulary, and an apparatus for implementing the method.

5

BACKGROUND OF THE INVENTION

The invention relates to a method for operating a speech recognition arrangement. In particular, the invention relates to a method as recited in the preamble of Claim 1. At present, various appliances are becoming voice-controlled. The operative outputting may imply executing a command, asking for further speech input, or having another result that is appropriate for the appliance in question.

Now, many present-day dedicated speech recognition circuits use only a small active vocabulary, which smallness feature brings about the advantage of fast and accurate recognition. An associated disadvantage is however that only commands and similar items within a limited set thereof can ever be recognized. Reloading of a different vocabulary into the arrangement is a relatively slow and tedious process that, even if successful, would slow down an interactive recognition process appreciably, in particular, because the presentation of a next speech item would have to wait until recognition of all preceding speech items were accomplished.

The present inventors have recognized that recording of the sound as received is often not the bottleneck, and that a hierarchical recognition procedure would allow to combine a relatively large overall vocabulary with the possibility for allowing uninterrupted continuous speech inputting. Moreover, they have found that users may well be saying there right words, but are prone to present an incorrect sequence of speech items. In fact, irrelevant items may be inserted, the sequence of the words may be inverted with respect to the correct one, and so on. In consequence, the inventors have come up with the idea that the recognizing of certain speech items might be delayed or abandoned, in favor of the recognizing of other speech items as received, to so improve flexibility by means of simple and inexpensive measures.

The inventors have also recognized that the calling up of a next speech item may be an order of magnitude faster than the calling up of a next partial vocabulary, in as much as such vocabulary may contain a much larger number of items-for-recognition.

5

SUMMARY TO THE INVENTION

In consequence, amongst other things, it is an object of the invention to allow uninterrupted speech inputting whilst effectively using a larger vocabulary, of which instantaneously only a smaller part is active through a hierarchical selection in said larger

10 vocabulary, and furthermore to allow for delaying or abandoning the recognizing of certain problem items. In fact, according to the present invention, a successfully recognized word or other item would tend to point both to another speech item as well as to an associated vocabulary, whereas an unsuccessfully recognized word would tend to point to another word whilst in contradistinction maintaining the vocabulary.

15 Now therefore, according to one of its aspects the invention is characterized according to the characterizing part of Claim 1.

The invention also relates to a apparatus arranged for implementing a method as claimed in Claim 1. Further advantageous aspects of the invention are recited in dependent Claims.

20

BRIEF DESCRIPTION OF THE DRAWING

These and further aspects and advantages of the invention will be discussed more in detail hereinafter with reference to the disclosure of preferred embodiments, and in particular

25 with reference to the appended Figures that show:

Figure 1, a block diagram of a speech-enhanced device;

Figure 2, a more detailed block diagram of the speech recognition facility;

Figure 3, a flow chart of the hierarchical speech recognition;

Figure 4, a diagram of an exemplary command structure.

30

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Figure 1 illustrates a block diagram of a speech-enhanced device, such as for effecting a consumer audio/video functionality. However, other environments, such as games,

appliances, motor vehicles, telephone sets, etcetera are equally viable. In the embodiment, the user functionality implies television set 24, tuner 20, antenna 22, CD player 28, and loudspeaker subsystem 30. These subsystems are controlled and interconnected by a central control facility 26. Control commands will emanate from facility 26, such as being originated by user button

5   actuation or remote control, not shown, or rather, according to the present invention, emanating from a speech recognition facility housed in subsystem 32 that bidirectionally interfaces to control facility 26. Speech-to-be-recognized is received on microphone 34 as a stream that may continuous, and which stream can be recognized in recognizer subsystem 36 through selectively accessing and addressing speech memory 38. If applicable, the speech channel may be

10   bidirectional viz à viz a user person, but this is not a restriction.

Figure 2 illustrates a more detailed block diagram of a speech recognition facility embodiment. The recognition facility 40 receives the speech through microphone 42, whereupon the speech is preprocessed to suppress background and other unwanted noise, and to execute filtering and similar preliminary operations. Block 46 executes delimiting of the speech

15   to break it down into separate words or other items that can be recognized separately. Another procedure is to store unpreprocessed speech, to be processed before recognizing. In block 48 the string of words and/or other items is stored in separately accessible memory locations to allow selecting of a particular recognizable such item. In block 50 the items are successively called in and subjected to a recognition process, as based on an actual and generally partial vocabulary

20   extracted from storage subsystem 54. The outcome of the recognition in block 50 can call up a next item for recognition, and also call up an associated partial vocabulary that is based on the identity of the recognized item from block 50. However, if the item is not recognized, the system may retain the actual vocabulary, but go to a new item. Most often, this will be one that succeeds the unrecognized item in time. If applicable as based on the status of the recognition

25   process, a command for the remainder of the system will be sent out on line 52 that may be connected to block 26 in Figure 1. On the other hand, certain results can make the apparatus signal a failure to the user. The apparatus may then ask for a whole or partial new speech input (for example if one word is missing or a parameter value is out-of-range), signal that the command cannot be executed (for example if an associated subsystem has been taken away), or

30   other.

Figure 3 illustrates a flow chart of the hierarchical speech recognition of the present invention. In block 60, the process is started though claiming and assigning the necessary hardware and software facilities. In block 62 a first partial vocabulary is called up. In block 64 the first word or similar recognizable item is called up from memory 48 in Figure 2. In

block 66 the speech item is subjected to the recognition processing. In block 68 the system

detects whether a successful recognition has indeed been effected. If positive, in block 70, the

result is retained. In block 72, the system detects whether the amount of items recognized is

sufficient for operative outputting. If no, in block 74 the new partial vocabulary is selected. In

5   block 76, the new speech item is selected. Then, the system goes to block 62 again. Eventually,

the answer in block 72 may become positive. Then the system goes to block 78 for the operative

outputting, and the operation stops in block 80.

        If negative in block 68, the system goes to block 82 and looks whether the item

was the last to be taken into consideration. This may be the last item in time, or rather the last

10   item that could help in finding an ultimately good result. If not the last, the system in block 84

selects the next appropriate speech item and reverts to block 64. Often, this will be the next item

in the progress of time, but under certain circumstances, this stepping will be counter to the

progress of time. If the item causing the failed recognition was the last indeed, in block 86 the

system detects whether additional information from the user could help, such as the pressing of a

15   button or the presenting of a particular item (not necessary the most recently recognition sought

for). If yes, the system in block 88 asks for such item and in block 90 checks whether the input

was the one sought for, and the recognizing has attained its goal (just as in block 72). If positive,

the system goes again to block 78. If negative in block 90 or in block 86, the additional

information had been judged as insufficient anyway, the system goes to block 92, wherein

20   failure is signaled, such as by asking for a complete new utterance from the user. Then the

system goes to block 80.

        The system could feature a much more sophisticated retry organization, such as

in that the negative outputs from blocks 86, 90, could be considered more fully. For example,

after block 86 a new partial vocabulary could be chosen, such under the assumption that one of

25   the recognized speech items had in fact not been intended by the user person, who had then

effectively proceeded with another series of speech items. Only if the recognizing were to stall a

second time, the system would then signal the failure of block 92. Various other tactics under

the strategy of the present invention have been ignored for the sake of brevity.

        Figure 4 illustrates a diagram of an exemplary command structure, in particular

30   for use with a multi-device audio/video environment. The first block represents an initial

vocabulary to address the appropriate device, such as the TV, the CD player, or the Tuner. The

next column of blocks represents for each address an appropriate set of Actions, that for a TV

set could allow selecting Teletext, a particular Channel, various Settings, and the like. Similarly,

for a CD-player, the next column of blocks may specify Play, Change to a next CD, Adjust

Volume, etcetera. Also, a Tuner entry has been shown in the first column, but it should be recognized that various other devices could be present, such as a CD-writer, a Screen Printer, various Household Appliances, etcetera. Also in the next column of blocks, not all feasible entries have been shown.

5   Upon specifying the Action in question, the third column of blocks will therefor specify an Action Parameter, such as selecting a Sports section of Teletext, for Channel a particular Channel, for Settings a particular aspect of the Settings that is wanted to be adjusted. Similar choices have been implemented for the CD player (second block in the second column) and for the Tuner. Of course, the hierarchy may even be deeper than the three successive levels

10   shown in the embodiment.

          Now, various methods for switching among the respective vocabularies are feasible. A first one is dependent upon the sequence of recognition: here the choice depends on the speed of "recognizing recorded items or sections" and the attainable speed of the "vocabulary switching". In the above context, the switching may be based on *word spotting*.

15   Here, every section of a spoken sentence is tested versus the active sub-vocabulary. Once a word is recognized, the associated vocabulary on the next lower hierarchical level will be loaded. Another strategy will recognize *words in a spoken sequence*. Here, the first word spoken will be tested versus all vocabularies, and then the next word, whilst presenting the vocabularies in a logical sequence. Once a word will have been found, the system loads a vocabulary that is

20   logically related to the result of the recognition. For example, when the system finds the phrase "sky radio", it knows that it has to use the radio-related procedure. For example, it can already start the tuner, whilst in parallel therewith proceeding with the recognition, and checking subsequent speech items to verify that the actual user person wants to more specifically deal with the music, such as recording it or otherwise.

25   The structuring of the vocabularies can be executed in various different ways. A first one is the rigid organization shown in Figure 1. A second organization has been termed amorphous. This may be used when various vocabularies share a particular command or other speech item. This allows to create combinations of speech items that match with the various different commands that users may present at a particular stage in time of operating a particular

30   device. For example, the speech as received may contain various speech items that are by themselves superfluous or even contraproductive. The recognition of a combination of a plurality of speech items allows to phase-out such irrelevant speech items. The speech recognition may allow for completely hands-free operating of a particular device. However, such may combine with various non-speech elements of a user interface organization, such as a

mechanical action by the user in the form of mechanically inserting a tape or disk, the actuation of a control button, or a stylus entry. Also, self-reliant operation of the device to be controlled can influence the effect of the speech recognition, such as the attaining of a tape's end, which would render the changing of the volume control superfluous.